A SHORT-LATENCY UNIT SELECTION METHOD WITH REDUNDANT SEARCH FOR CONCATENATIVE SPEECH SYNTHESIS

Nobuyuki Nishizawa[†] *and Hisashi Kawai*[†‡]

[†]ATR Spoken Language Communication Research Laboratories, Japan
[‡]KDDI R&D Laboratories Inc., Japan
{nobuyuki.nishizawa, hisashi.kawai}@atr.jp

ABSTRACT

A new method for short-latency unit selection is proposed. For prompt response in concatenative speech synthesis systems with large unit databases, waveforms should be output before all speech segment units of an utterance are determined. For that purpose, short-latency unit selection algorithms were introduced in our previous study. However, the short-latency unit selection may cause degradation of quality because units that consist of the optimal unit sequence may be pruned by forcible unit determination on the search. In the proposed method, the degradation of quality is suppressed by redundantly expanded hypotheses based on N-best search. The results of unit selection experiments in a practical configuration indicate that the proposed method is superior to the conventional DP search method when latency in unit selection is set to be short.

1. INTRODUCTION

In waveform concatenative speech synthesis, the quality of speech sounds depends on the suitability of each waveform segment. If more waveform segment units are available, more suitable units can be found in unit selection. For this reason, large-scale unit databases are often used in concatenative speech synthesis.

This suitability includes not only similarity between a synthesis target and a selected unit but also the smoothness between neighboring units. Therefore, combinations of units should be considered in unit selection. Since computational effort in the search for the optimal unit sequence is proportional to the square of the number of possible units even in dynamic programming (DP)-based search, as a larger unit database is used for high-quality sounds, unit selection becomes much slower. However, in many applications, such as dialog systems, prompt response in speech synthesis is demanded. To achieve prompt response, two approaches are mainly considered: One is to improve the throughput, and the other is to reduce the latency in unit selection.

In using the latter approach, we previously introduced a short-latency unit selection algorithm, which was called a short delay unit selection algorithm[1]. In short-latency unit selection, units are forcibly determined before all hypotheses in search are expanded. Since unit possibilities that consist of the optimal unit sequence can be removed by the forcible unit determination on the search, the quality of sounds may be degraded by the short-latency unit selection. In that case, redundant hypothesis expansion compared with hypothesis expansion in complete DP search can suppress degradation. This is because good solutions in the redundant hypotheses may be left as possibilities after forcible unit determination, although the solutions may not be the optimum. However, in the previous study, only full hypothesis expansion was examined as a redundant hypothesis expansion technique. This search method was an extreme methodology for short-latency unit selection with beam search in practical configurations because the method was too sensitive to pruning by beam search. Therefore, in this study, a new algorithm based on DP search with N-best hypothesis expansion is proposed. Because the algorithm includes search algorithms based on the conventional DP and the full hypothesis expansion in our previous study as special cases, the new algorithm can be considered a generalization.

In this paper, the new algorithm is discussed. The rest of the paper is structured as follows: For later discussion, section 2 describes designs of cost functions for unit selection in our speech synthesizer. Section 3 explains a unit determination algorithm in short-latency unit selection. Then, in section 4, the new hypothesis expansion method based on N-best search is introduced. Evaluations of the new algorithm are given in section 5. Finally, section 6 concludes the paper.

2. COST FUNCTION FOR UNIT SELECTION

For speech synthesis, units in a database are selected based on the minimization of a criterion, which is often called *cost*. In our configuration, the integrated cost function C corresponds to the degradation of naturalness in unit sequences $\{u_i\}$. C is given by a recurrence formula:

$$\begin{aligned} C(0) &= w_T \cdot C_T(u_0, t_0) \\ C(i) &= C(i-1) \\ &\quad + w_C \cdot C_C(u_{i-1}, u_i) \\ &\quad + w_T \cdot C_T(u_i, t_i) \quad (i \geq 1) \end{aligned} \tag{1}$$

where C_T, C_C and t_i denote target cost, concatenation cost, and target information at time i, respectively. Also, w_T and w_C denote the weight coefficients of the target cost and the concatenation cost, respectively.

The target cost function C_T represents degradation of naturalness caused by the difference between a target and a selected unit in a mismatch of the phonetic environment, phone duration, log F_0 (fundamental frequency), and MFCC (mel-frequency cepstral coefficients). In our TTS (Text-to-speech) system, which is named XIMERA[2], all of these features except the phonetic environment are predicted by techniques of HMM-based speech synthesis[3].

On the other hand, the concatenation cost function C_C represents the degradation of naturalness caused by discontinuity at the unit boundary in F_0 and MFCC.

For higher-quality sounds, the target cost function, the concatenation cost function, and the weight coefficients in our synthesizer were estimated from the results of perceptual experiments[4].

3. SHORT-LATENCY UNIT SELECTION

3.1. Pipelined processing for short latency

In conventional unit selection, units were selected based on the minimization of the integrated cost of one utterance. Therefore, all possible units of the utterance must be evaluated before the determination of any unit of the utterance. Consequently, speech output is delayed, at least for the unit selection of one utterance. For prompt response, the evaluation must be several times faster than real time. However, if a large unit database were used, such rapid processing would be impossible.

For prompt response, the unit at the beginning of the utterance should be determined before the evaluation for all of the units is finished. For this purpose, a short-latency unit selection algorithm was introduced in our previous study. Figure 1 schematically shows an example of the processing order of procedures in the short-latency unit selection. In this example, the processes for the search of a unit sequence, waveform concatenation, and playback of waveforms are parallelized for shorter latency. In contrast to the conventional unit selection, only the first and second unit possibilities, not the first to the fifth unit possibilities, are considered in determination of the first unit. With this method, pipelined processing for shorter latency is achieved.

3.2. Determination of units

In the following, unit selection algorithms are discussed for search trees. In this study, only breadth-first search is considered for DP-like search algorithms. In breadth-first search trees, distances from the root node to all leaf nodes are equal. In the conventional DP-based search, search trees are expressed as breadth-first search trees where each unit at each time is linked to only one unit at the previous time. Simultaneously, beam search for search space restriction is also considered because a truly full search is impractical even in the DP search. Beam search in this study is implemented by restricting the number of kept hypothesis sequences. The limit number of the kept sequences is called the *beam width*.

In the short-latency unit selection, units are selected based on the minimization of integrated cost from the root node to the leaf nodes of an incomplete tree, not the leaf nodes of a complete tree. Because the tree depths at all leaf nodes are equal to each other in breadth-first search, integrated costs at the leaves can be simply compared for unit selection. To build a deeper search tree, only the first unit in the best hypothesis sequence is determined, and the determination is deferred as long as possible. As the depth of the search tree grows deeper, a better unit sequence as a whole will be selected. When a unit at each time is determined, all hypothesis sequences not including the determined unit become useless sequences for the search. Therefore, those sequences should be removed from the search tree. This operation corresponds to pruning all branches to sibling nodes of the determined node.

Figure 2 shows an example of a search tree in short-latency unit selection where the beam width w is 4 and hypotheses from $i = 1$ to $i = 4$ have already been expanded. The name at each node, such as a1 or b1, denotes a unit. Note that there are multiple nodes corresponding to the same unit because the figure is not an example of the conventional DP search. Using the beam width limit, some branches have already been removed. If the unit sequence that has the minimal integrated cost is R-a1-b2-c4-d2, unit b2 is determined as a selected unit at $i = 2$, because the unit is included in the optimal sequence. On the other hand, units c4 and d2 are not determined at that time. This is because subsequence b2-c2-d3 may be considered more suitable than subsequence b2-c4-d2 after hypotheses later than $i = 4$ are expanded. After unit b2 is selected, the other sibling nodes

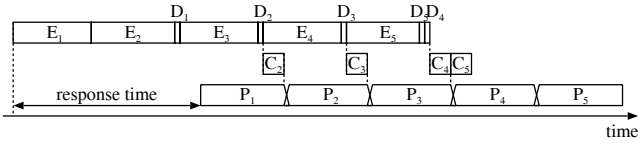

Fig. 1. An example of the processing order of procedures in short-latency unit selection. E_i, D_i, C_i, and P_i denote the expansion of hypotheses for the i-th unit, determination of the i-th unit, waveform concatenation between units of the (i−1)-th and i-th units, and playback of the waveform segment of the i-th unit, respectively

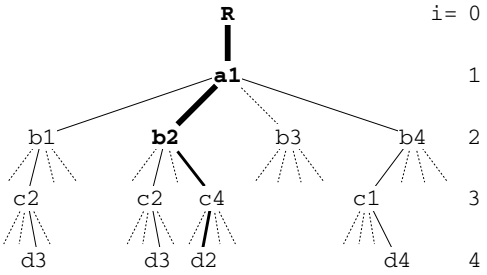

Fig. 2. An example of a search tree in short-latency unit selection. The thick bold, thin bold, solid, and dotted lines denote the determined branch, a branch that consists of the optimal unit sequence, a kept branch, and a removed branch by beam width limit, respectively.

b1 and b4, and their children are removed; in the example, unit b3 has already been removed due to the restriction on the beam width.

3.3. Requirement in the throughput

In practice, the time for unit selection must be at least shorter than real time for continuous speech output. When a large unit database is used, unit selection may still be slow even if the beam width is set to 1 as the minimal value. If a database is built from a 100-hour corpus, the number of possible units at a given time can number in the tens of thousands. Therefore, the number of possible units should be reduced before hypothesis expansion. This technique is called *pre-selection*. In this study, the reduction of units is performed by using the target cost. Consequently, units far from the target will be removed. The number of units after this reduction is called *pre-selection width*. Pre-selection was already adopted for CHATR[5], which is ATR's previous concatenative speech synthesizer.

In this study, it is supposed that the speed of unit selection is controlled by the beam width limit and the pre-selection width limit because the computation effort for unit selection is mainly affected by these.

4. SEARCH METHODS

4.1. Search based on the conventional DP search method

In the conventional unit selection, DP-based search, which is often called *Viterbi search*, is widely used. In the DP search, for each unit at a given time, only the best combination of expanded sequences at the previous time and the unit is expanded as one of new hypothesis sequences at the given time. This is because other combinations are not parts of the optimal unit sequence. Consequently, in search trees of the DP search, there is no duplicate subsequence. This efficiency is a major reason that the DP search is widely used.

By contrast, in short-latency unit selection, valuable subsequences may be removed from the tree by forcible unit determination. This is due to difference in criteria between unit determination and hypothesis expansion. Therefore, in short-latency unit selection, this search method is DP-like search, not DP search, correctly. Although re-expansion of hypotheses from the determined unit is necessary for strict DP search from the determined unit, the re-expansion is impractical in the short-latency unit selection because usable time for the re-expansion is not enough to rebuild a deep tree; the usable time is equal to duration of the last determined unit.

Figure 3 is an example of the DP search. The figure shows that there is no duplicate subsequence in the search tree. However, if unit b2 is forcibly determined for a unit at $i = 2$, three possible sequences R-a1-b1-c2-d1, R-a1-b1-c2-d3, and R-a1-b4-c1-d4 in total four kept sequences are removed. Even if sequence b2-c2-d3 is the secondary sequence from b2, it is not considered.

4.2. Redundant search based on N-best search

In contrast to the conventional DP search, if two or more redundant hypotheses corresponding to the same unit at each time are expanded, probability that the unit is removed by the forcible unit determination will decrease. Therefore, a new method with redundant hypothesis expansion is introduced. In this study, as an extension of the conventional DP search method, N-best hypotheses expansion is adapted to the DP search. In the method, not only the best hypothesis sequence but also N-best hypothesis sequences are expanded in hypothesis expansion of the DP search. If no pruning by beam search or the short-latency unit selection is done, the N-best solutions can be obtained by this method, i.e., the method corresponds to DP-based N-best search in the conventional unit selection. Therefore, the method is called *redundant search based on N-best search*. Although the number of expanded hypothesis sequence is production of pre-selection width k and N of the N-best search, the expanded hypotheses are reduced to beam width w by beam search. Therefore, the computational effort for the hypothesis expansion is less relevant to the value of N, and is mainly proportional to production of k and beam width w, which is equal to the number of combinations of possible units at a certain time and kept possible sequences at the previous time. In other words, computational effort of the proposed method is almost same as that of the conventional search method.

When the value of N is equal to 1, this search method is equivalent to the conventional DP search. On the other hand, when the value of N is equal to beam width w, the method is the same as the full expansion search method in our previous study because all possible hypotheses are expanded. Therefore, the DP search and the full expansion search are special cases of the redundant search.

However, in the same w, the performance of the redundant search may be inferior to that of the conventional DP-based search because in search for the best sequence, the redundant search is equivalent to narrowing the beam width if influence of pruning by the forcible unit selection is not considered. Therefore, the redundant search is useful only when the degradation of performance by pruning is more critical than that by narrowing beam width.

Figure 4 is an example of the redundant search based on 2-best search where $w = 4$. In this example, it is supposed that subsequence c2-d3 is a good subsequence. Compared to the case in the conventional DP search, there is a higher probability that subsequence c2-d3 is kept in the search tree. Even if the optimal sequence is R-a1-b2-c4-d2 just after possible units at $i = 4$ are expanded, the subsequence is kept. Generally, a sequence that includes good subsequences may be good although the sequence may be not the optimum. Therefore, sequences like R-a1-b2-c2-d3, which includes

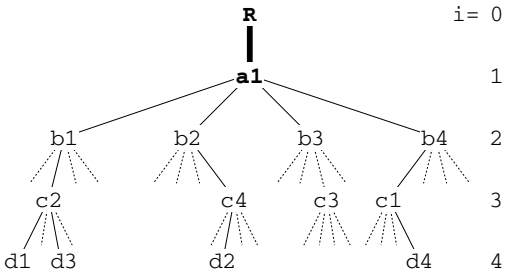

Fig. 3. An example of a search tree in search based on the conventional DP search. There is no duplicate subsequence.

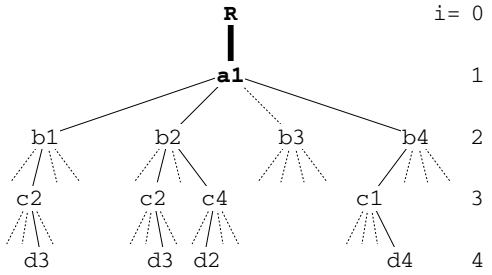

Fig. 4. An example of a search tree in the redundant search method based on 2-best search. Redundant expansion of possible subsequences (c2-d3 in this example) is permitted.

subsequence c2-d3, tend to be kept even in beam search. However, variety of subsequences is inferior to that in the conventional DP search.

5. EXPERIMENTS

To evaluate short-latency unit selection with the redundant search, integrated costs of selected unit sequences were investigated for various pre-selection widths, beam width limits, and search depth limits. The search depth means the total target duration of units that are not determined yet, in a hypothesis sequence. In the experiments, in order to examine the relationship between search depth and the result of unit selection, it is supposed that the time for unit selection is always equal to real time.

The unit database used in this evaluation was built from a Japanese corpus of approximately 59 hours pronounced by a female speaker. Each unit in the database corresponds to a half-phoneme for vowels and unvoiced fricatives, or a phoneme for the other consonants. Half-phoneme units are adopted for waveform concatenation at the center of the phoneme. This is because waveform concatenation at the center of a phoneme is often easy in Japanese[6]. The synthetic targets are 53 sentences of J set in ATR's 503-sentence corpus[7]. The target information was extracted from utterances pronounced by the same speaker.

In the first experiment, the product of k and w is always fixed to 100,000. This configuration is estimated by a preliminary experiment, and can practically be conducted in recent PCs. Note that N of the N-best search is virtually irrelevant to the computational effort for unit selection especially when the value of N is small. Figure 5 shows the mean integrated cost per unit of selected units by short-latency unit selection with the redundant search. The figure indicates that the optimal N is not equal to 1, i.e., the redundant search is superior to the conventional DP search when the value of

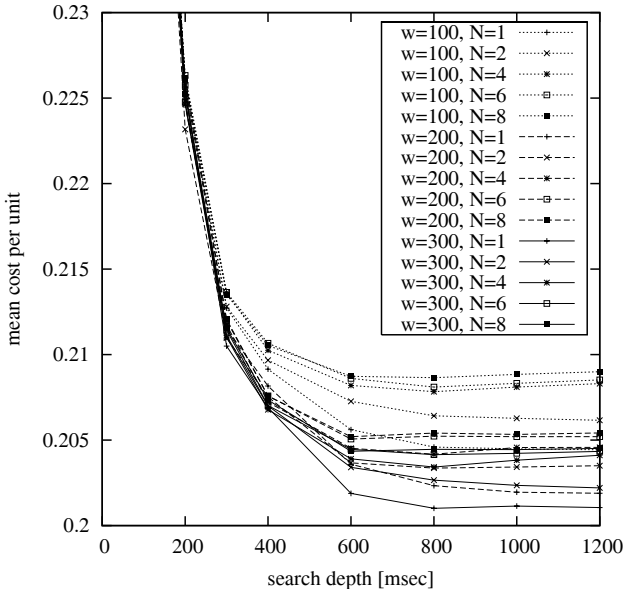

Fig. 5. Mean integrated cost per unit of selected units by short-latency unit selection with redundant search where products of the beam width and the pre-selection width are fixed to 100,000. w and N denotes the limit of beam width and the value of N in the N-best search, respectively.

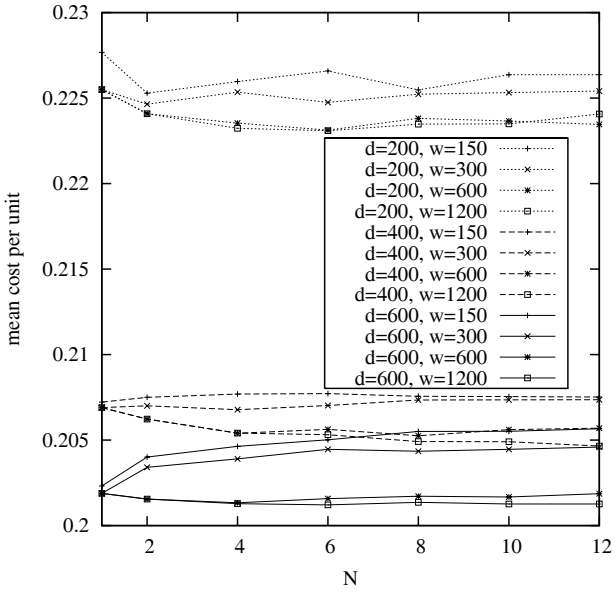

Fig. 6. Mean integrated cost per unit of selected units by short-latency unit selection with redundant search where the pre-selection width k is fixed to 333. d, w, and N denotes the search depth limit, the beam width limit, and the value of N in the N-best search, respectively.

the search depth limit is shorter than 300 milliseconds. However, the difference between the optimal results of the proposed method and the results when $N = 1$, which corresponds to the conventional DP search method, is slight. The result implies that suppression of the degradation of quality caused by the forcible unit determination is difficult only by the redundant search especially when search depth is extremely shallow.

In the second experiment, only the pre-selection width k is fixed. When the k is constant, as the beam width w become larger, the optimal N will also becomes larger. In this experiment, k is fixed to 333, because in another preliminary experiment, where w was fixed, the cost of selected unit sequences almost converged when k was larger than 333. Figure 6 shows results of the experiment. The figure indicates that the optimal N is larger than 1, and when w is large. In other words, the redundant search is effective when the bean width is wide, even where the search depth is over 300 milliseconds. For the present, configurations where w is larger than 300 are impractical in PCs because the time for unit selection is exceeds real time. However, the pre-selection algorithm in this study is too simple. Therefore, for larger w, smaller k may be possible without degradation of cost by more effective pre-selection methods, for example, using a priori knowledge. In another case, faster computers will make these configurations practical in the near future.

6. CONCLUSION

A redundant search method based on N-best search was introduced to short-latency unit selection. Degradation of quality is suppressed by redundantly expanded hypotheses. In the proposed method, redundancy of the hypothesis expansion is controlled by the N-best search. To evaluate the proposed method, unit selection experiments were conducted. The results of the experiments indicated that the redundant search is superior to DP-based conventional search without redundancy when latency in unit selection is set to be short and/or beam width is set to be wide.

Acknowledgements: This research was supported in part by the National Institute of Information and Communications Technology.

7. REFERENCES

[1] N. Nishizawa, H. Kawai, "Using a Depth-Restricted Search to Reduce Delays in Unit Selection," Proc. ICSLP2004, Jeju Island, Korea, vol. 2, pp. 1209–1212, Oct. 2004.

[2] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, K. Tokuda, "XIMERA: A New TTS from ATR Based on Corpus-Based Technologies," Proc. 5th ISCA Speech Synthesis Workshop, pp. 179–184, Pittsburgh, Pennsylvania, U.S.A., June 2004.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis," Proc. ICASSP 2000, Istanbul, Turkey, vol.3, pp.1315–1318, Jun. 2000.

[4] T. Toda, H. Kawai, M. Tsuzaki, "Optimizing Integrated Cost Function for Segment Selection in Concatenative Speech Synthesis Based on Perceptual Evaluations," Proc. EUROSPEECH, Geneva, Switzerland, pp. 297–300, Sept. 2003.

[5] A. Hunt, A. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," Proc ICASSP 96, vol. 1, pp 373–376, Atlanta, Georgia, U.S.A., May 1996.

[6] T. Toda, H. Kawai, M. Tsuzaki, K. Shikano, "Unit Selection Algorithm for Japanese Speech Synthesis Based on Both Phoneme Unit and Diphone Unit," Proc. ICASSP2002, pp. 3088–3091, Orlando, Florida, U.S.A., May 2002.

[7] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, Speech Database User's Manual, ATR Interpreting Telephony Research Laboratories Technical Report, TR-I-0166, Japan, Aug. 1990 (in Japanese).